Supplement

S1. PathfinderTURB algorithm and data

The *S* signal in eq.1, is used by PathfinderTURB to retrieve the CBL and the TCAL at PAY (resp. KSE) with a time resolution of 1 minute (resp. 2 minutes) and a range resolution of 30 m (resp. 45 m). *S* is also smoothed in space and time (2-D) using a Gaussian filter with FWMH equal to 5 bins in range corresponding to 150 m (resp. 225 m) and 5 steps in time corresponding to 5 minutes (resp. 10 minutes) and a standard deviation equal to 1.1 bins in range and 1.1 steps in time. Then smoothed with the 2-D Perona-Malik anisotropic diffusion method (Perona and Malik, 1990, as suggested in Haeffelin et al., 2012), based on 15 iterations and a conduction controlling coefficient of 2500. Values less than 1000 are set to 1000 and then the log10 is taken. The corresponding gradient field is calculated with central differences of this log10, smoothed RCS.

By taking the central differences of the log, we can relate it easily to the vertical ratio of RCS, which simplifies the setting of some parameters of the algorithm:

$$grad\left(log10\left(RCS(R_i,t_j)\right)\right) = \frac{\left(log10\left(RCS(R_{i+1},t_j)\right) - log10\left(RCS(R_{i-1},t_j)\right)\right)}{R_{i+1}-R_{i-1}}$$
$$= \frac{1}{R_{i+1}-R_{i-1}} log10\left(\frac{RCS(R_{i+1})}{RCS(R_{i-1})}\right)$$

In Section 3.2, the criteria for the PathfinderTURB to restrict *S* and ∇S in range and time by applying a set of physically based restrictions were described. Applying these restrictions is an important step of the algorithm. We give next further details on the restrictions for strong negative and positive gradients, as well as on the calculation of the atmospheric variance and of the TCAL. The last section describes the guidelines and data used for the human expert MLH retrieval during the PathfinderTURB validation in Payerne.

S1.1 Strong negative gradients

During the early morning period, the threshold is set to *log10(0.85)/(2dR)*, which is the value of the derivative when the signal drops more than 15%. Else it is set to *log10(0.75)/(2dR)*, (signal drop of more than 25%). The signal is looked from 250m up.

In case there is a PBL cloud above, the limit is set to the PBL cloud limit.

S1.2 Strong positive gradients

During the early morning period, the threshold is set to *log10(0.95)/2dR*, which is the value of the derivative when the signal gains more than 5%. Else it is set to *log10(0.85)/2dR*, (signal gain of more than 15%). The signal is looked from 250m up.

In case there is a PBL cloud above, the limit is set to the PBL cloud limit.

If there is strong negative gradient less than 300m above a strong positive gradient, the limit is set at the strong negative gradient's altitude.

The strong negative gradient, strong positive gradient and lower altitude restrictions are all relaxed over 5 minutes, i.e. for a given time step, the highest restriction in a 5-minutes window centred on that time step is taken.

S2. Calculation of the first transition to turbulence, atmospheric variance and total variance

For MLH retrieval algorithms, it is interesting to look at the variability (in time) of the RCS, because the backscatter signal displays usually high temporal fluctuations at the top of the MLH. However, when computing the variance of a RCS time series, meso-scale to micro-scale and noisy fluctuations are taken into account. We can use spectral analysis (numerically using FFT) to try to recover the fluctuations solely due to aerosol number density fluctuations, but also to estimate whether we are in a fully developed turbulent regime, by looking at the spectral power decay. This gives us then, at each range and each time step, an estimation of the atmospheric-only variability and of turbulence presence.

Let us be given a 1h-long time series of the RCS, centered on a given time step at a given altitude. The method, based on (Pal et al., 2013), goes as follows:

- Despike the time series recursively with a sliding histogram as in (Lenschow et al., 2000). This step is used to remove extreme values due e.g. from clouds.
- Detrend with a polynomial regression of order 2 (i.e. substract the fit from the timeseries).
- Calculate the FFT.
- Reduce the influence of meso-scale variability (periods of 30 min and more) with a high-pass filter applied on the FFT coefficients. We use a raised cosine filter (high-pass cutoff f_{cutoff} = 1/1800 Hz and roll-off factor 0.5).
- Integrate the power of the high-pass filtered FFT between −0.75 f_{Nyquist} and 0.75 f_{Nyquist}. This gives us the atmospheric variance var_{atm}.
- Integrate the power of the high-pass filtered FFT everywhere. This gives us the total variance var_{tot}.
- Perform a linear fit of the power spectrum, in log scale, between the 10-minutes frequency and the frequency corresponding to 0.75 f_{Nyquist}, and retrieve the slope. Calculate then the relative error err_{rel} of the slope value compared to the theoretical value of −5/3 (Kolmogorov's power law for the inertial subrange) and apply the power 4 to this relative error.

var_{atm} is an hourly approximation of the atmospheric variance. We divide it by the total variance var_{tot} value and apply the power 4 to have a larger span of the values between 0 and 1, because this ratio of atmospheric vs. total variance is about 0.75 for a signal with fluctuations dominated by noise and is close to 1 for a signal with fluctuations dominated by aerosol number density fluctuations. The resulting value, $atmvar = (var_{atm} / var_{tot})^4$, is a measure of the atmospheric variability.

In mixing and entrainment zones and periods, the -5/3 scaling law of turbulence for the inertial subrange can be observed. The relative error is then a good proxy to see if the region observed is in a turbulent regime or not. Again, we apply the power 4 to have a larger span of the values between 0 and 1 and so better distinguish between less and more turbulent zones. The resulting value, $turbfit = (1 - err_{rel})^4$, is a measure for the presence of turbulence.

S2.1 Implementation

From 00:00 to 24:00, in steps of 10 minutes (in order to reduce the computational time), the atmospheric variability *atmvar* is calculated, as the measure for turbulence presence *turbfit*, over periods of 1h centered on the time steps, for each altitude between the ground and the maximal climatological value set before.

The signal used here is the raw RCS from the ceilometer (*beta_raw)*, whose time and range resolutions are reduced to 1 min resp. 30 m. The Nyquist frequency f_{Nyquist} is then 2min and the first noise frequency 0.75 f_{Nyquist} is 2min40s.

The atmospheric variance *atmvar* and the relative error *turbfit* are first smoothed in range with a running mean of 11 points (i.e. over 300m), then both in range and time (2-D) using a Gaussian filter with FWMH equal to 5 bins in range corresponding to 150 m and 5 steps in time corresponding to 50 minutes and a standard deviation equal to 1.1 bins in range and 1.1 steps in time. Finally, they are interpolated on the time-and-range grid of the gradient field calculated before (i.e. (1min, 30m) in PAY, (2min, 45m) in KSE).

S2.2 Inferior limit for altitude from transition to turbulence

After the early morning period and before sunset, the altitude of the first transition to turbulence is calculated by applying a Haar-wavelet covariance transform on the profiles of *turbfit* and average it over multiple scales (60m to 240m in PAY, 90m to 360m in KSE), then we look at the first local minimum of this averaged transform.

If deemed necessary, the retrieval of the MLH could be made in an even more robust manner by varying the starting times and window widths, as described in the next subsection. However, doing so did not change much the results presented in the present study.

S3 Robust choice of MLH

The pathfinder has an inherent uncertainty, and its sources are due to the a-priori range restrictions, the gradient and variance calculations, the window width and the starting time. We consider here only the latter two uncertainty sources, and run the algorithm from sunrise to sunset with offsets of -30 to 30 min in steps of 10 min on the sunrise time, and window widths of 15 min to 45 min in steps of 5 min. This makes us run the algorithm 49 times, which increases the computational time to about 300 s in average. Parallelization of this step is straightforward such that, if needed, the computational time could be reduced easily on nowadays standard multicore machines.

To select the final MLH time series, we create a count matrix, which counts the number of MLH candidates at a given altitude and at a given time step and divides this number by the number of MLH candidates at this time step. This gives a value between 0 and 1. It is sometimes not enough to select the final MLH time series based on this count matrix, and so we multiply this matrix by a ratio matrix, which at each altitude and each time step calculates the ratio of the mean RCS 0m to 150m above and of the mean RCS 0m to 150m below the considered altitude. Negative ratios or ratios over 1 are set to 1. We subtract then the ratio from 1 and multiply the result by 2. This gives a value between 0 and 2, values close to 2 belonging to clouds, else the values are mostly below 1.

For each MLH time series, we add the values of the product of the count and ratio matrices along the time series path. The time series having the maximal value is then selected as final MLH time series.

S4. Calculation of the TCAL

The *S* signal in eq.1 has a time resolution of 2 minutes and is smoothed in space and time (2-D) using a Gaussian filter with FWMH equal to 5 bins in range and 5 steps in time and a standard deviation equal to 1.1 bins in range and 1.1 steps in time. The corresponding time resolution restriction and the 2-D Gaussian smoothing are applied to the detected noise level by using error propagation and ignoring non-diagonal covariance terms. The SNR is then calculated simply dividing *S* by the noise level.

The signal is considered too noisy when the SNR drops below 0.6745 or when *S* becomes artificially negative (which happens due to the manufacturer automatic background light subtraction procedure). This creates a binary mask the same size of *S* with the limitation that isolated points are not seen as noisy even if they have neighbouring noisy points and vice versa. Following (Biavati, 2014), image processing morphological operators, three erosions followed by 20 dilations, are used to make such isolated points disappear. Then, at each time step, the SNR mask is set to 0 at and above the first range bin above 600m where the mask is 0.

The discrimination of pure molecular signal from the aerosol signal is performed here by setting a threshold value on the profile of log10(*S*). Once we have determined the instrumental calibration constant using the procedure suggested by (Wiegner and Geiß, 2012) and after we have calculated the molecular profiles over one year with the formulas given in (Bucholtz, 1995) (Pressure and temperature from radiosounding), a threshold *Th* to discriminate aerosols from molecules can be derived as *Th*=4.625. We then define a binary mask for aerosols based on the threshold *Th*. Each profile of log10(*S*) is smoothed in range by applying a 11-point running mean at all ranges above the first range bin where the overlap function is larger than 0.05, and the aerosol mask is set to 0 at and above the first range bin where the running mean is lower than *Th*. The aerosol mask is also set to 0 at all ranges above the detected cloud base. Then, three erosions followed by 10 dilations are applied to the mask. For profiles where *S* was already smaller than *Th* at the ground, the mask is set to 0 at all ranges. This results in a non-molecular layer binary mask.

At each time step, the aerosol layer top is given by the first range bin where the product of the SNR mask and of the non-molecular layer mask is 0. The actual temporal resolution of the calculated aerosol layer tops is then decreased to 10 minutes (5 time steps), i.e. for each time step, the highest top calculated during 5 time steps centred on the current time step is selected. In case of low cloud, the aerosol layer top is not considered for the analysis.

S5. Human expert MLH retrieval in Payerne

The following guidelines were used:

- Each detection should lie on or be very close (within ±2 range bins) to an aerosol gradient.
- Determine if a manual detection is possible (i.e. is it possible to follow a gradient clearly for at least 30 min?):
    - In case of cloud: the detection must match the cloud base.
    - In case of fog or low stratus cloud: detection allowed when it starts to dissipate.
    - In case of precipitation: detection possible only during time intervals without precipitation of at least 2 hours.

- Meteorological and astronomical information should be taken into account in order to respect the following criteria:
    - The detection starts after sunrise and ends before sunset.
    - The CBL should start developing from the ground after sunrise, i.e. elevated layers shall not be selected during early morning stage.
    - Look for a CBL decay/breakdown when the vertical heat flux at the ground drops to about 0 and ΔT is a negative local minimum value at the end of the day, with ΔT calculated once per hour by taking the difference between two consecutive hourly means of the surface temperature measurement. If no decaying layer can be followed, no detection is to be provided.
- The expert selects as many points as it is necessary in order for a linear interpolation between two adjacent points to represent trustworthily the short-term evolution of the observed layer.

All information are taken into account (station measurements, MLH estimations from remote-sensing instruments and radio-sounding launches, synoptic conditions, etc.) and compared to the expert detection. In case of clear mistakes by the expert, the detection points should be edited accordingly.

S5.1 Data evaluation

The data were processed by either two or three experts in parallel:

- Three experts processed the 5, 10, 15, 20, 25 and 30 (if not February) of each month of the year 2014.
- Two experts processed the whole months of January, March, July and October 2014.